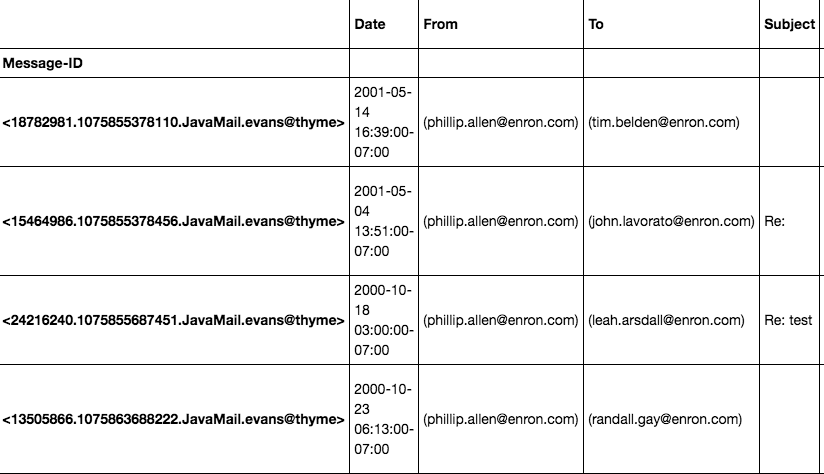
**CMPE258**

**Project- Email Auto-Compose**

**Samuel Yang, Yuanzhe Li, Yuhua He, Jia Ma, Ying Liu**

**Conclusion**

**In this group assignment, the business objective we are focusing on, is the "Smart Email Compose" project. During the data summarization, we are using the Enron email dataset, which has the features like this: Date, From, To, Subject, Content, and so on. We decided to do to main things on this project. First, we gather sentiment information to analyze if there polarity of each mail relates to the stock price chaingng of Enron Company in same period. Second, we will use RNN along with Long-Term-Short-Memory method to perform word generations, or text predictions.**

****

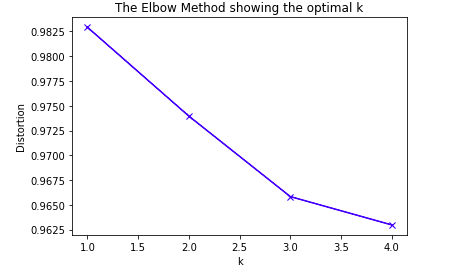
**Before reducing noisy data, we decided to perform data enrichment to up-sampling our data. We concluded that we would apply Latent Dichrilet Analysis(LDA) to categorize them to about ten topics. We will do Sentiment Analysis to the data added with the enriched data (message chatting) . The graph below is the enriched data.**



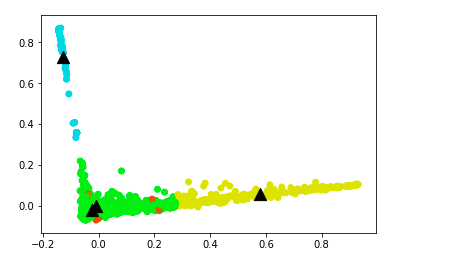
As for the data preprocessing, we use stemming and regular expressions to clear out stop words, titles, and other noisy data. It doesn’t have outliers and missing data. During the data transformation, we used LabelEncoder and bag of words method to help us to make K-means and GMM. We used elbow method to get the preferred number of clusters, which is “four”.

**Unsupervised model:**

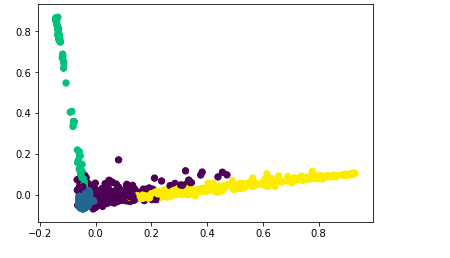
Elbow Method:



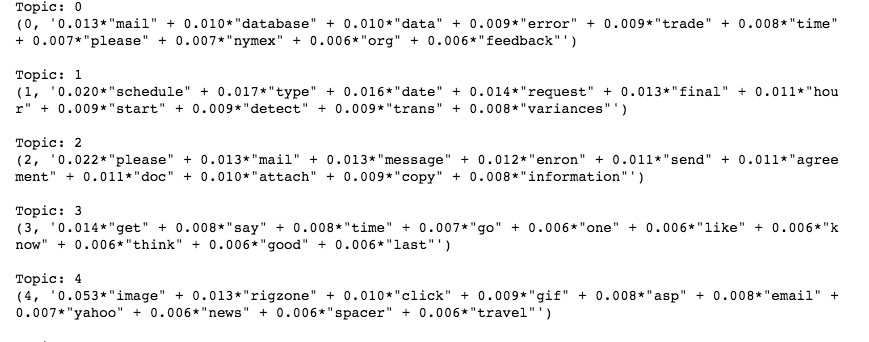
K-means



GMM:

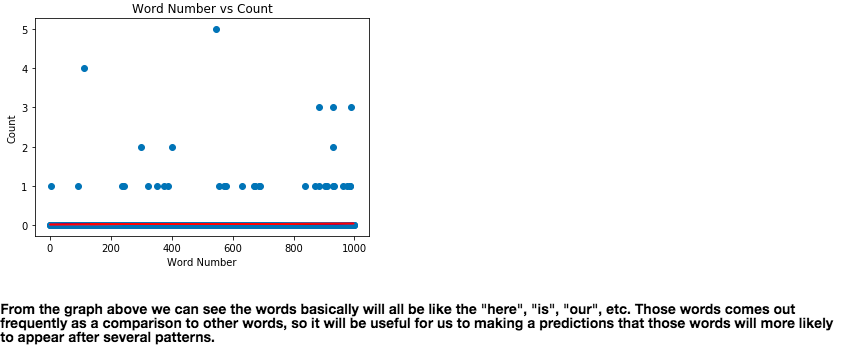


LDA:

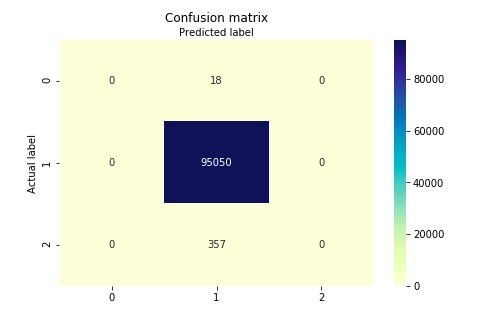


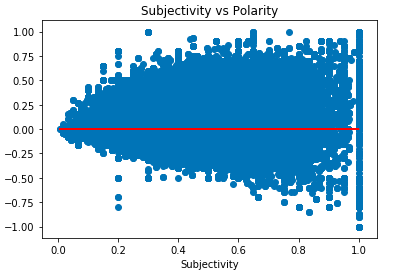
**Supervised Model:**

Linear Regression

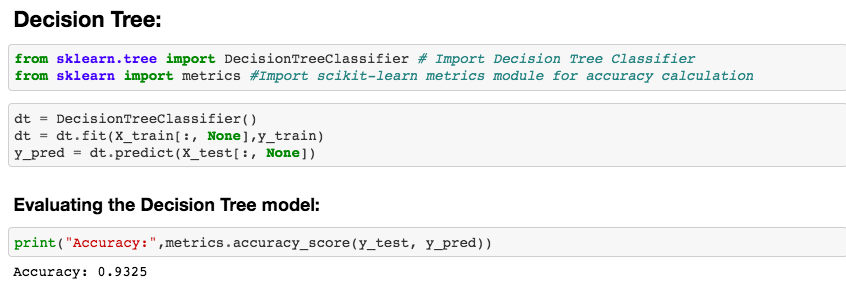


Logistic Regression and its confusion matrix:

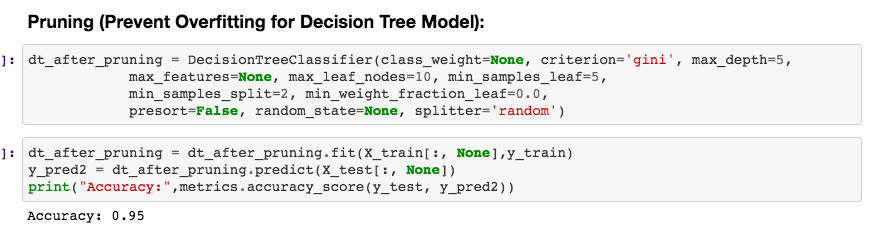




Decision Tree:



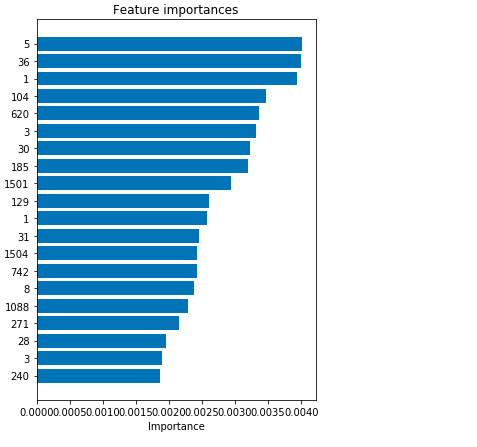
With Pruning:



**Embedded Model:**

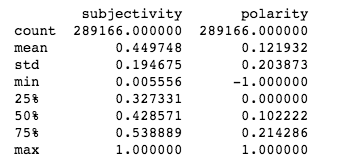
**Random Forest:**

**We uses RF to improve the accuracy and also prevent overfitting as well.**

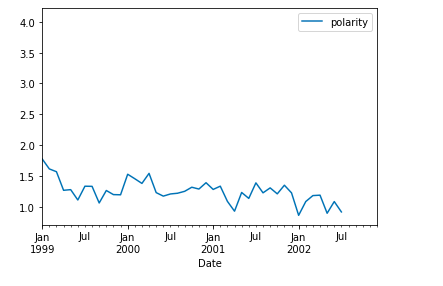


**So now we are going to see the correlations between the amount of emails, sentiments, and the share prices.**

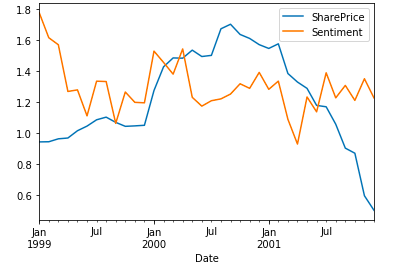
Sentiment Analysis:



Sentiment Distribution:



Stock Price vs Sentiment



From the Sentiment distribution and amount of emails sending in particular period, and adding the Share Price into these data. In conclusion, from all the data analyses above, there might be indications that during the Enron period there might be some unusual activities and emails corresponding to the change of the share price. When share price changes, the amount of emails in that period also goes up a lot. As a result, they might secretly have some businesses behind that we don’t know.

For the auto complete model. We trained a fairly model which able to generate a 50 words email, based on given Enron email data set. Training such a big model is time consuming, even with GPU acceleration, we it still cost us more than 7 hours. Our result is showing below(Defined in Copy\_of\_email\_genarator.ipynb):

This is the training and validation loss chart during training.

For the email generator(test\_model.ipynb):

As we can see here, give a random selected text slice from original email data set(sequence pattern), it can generate a new text base on given content. Like the seed text is about give a feedback ASAP, so the generative text is about how to make a response.